DOCUMENT RESUME

ED 106 354 TM 004 470

AUTHOR Enger, John M.; Whitney, Douglas R.
TITLE A Generalized Anova Model for Estimating the Reliability of Categorical Judgments.
PUB DATE [Apr 75]
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Analysis of Variance; *Classification; Measurement Techniques; *Models; Rating Scales; Reliability; Statistical Analysis; *Test Reliability

ABSTRACT

There are few existing or widely known measures of agreement applicable when data is nominal or categorical. Most such coefficients are applicable only when judges classify objects or subjects into a single category. A wider range of applications, including those where judges (1) place probabilities on subjects belonging to mutually exclusive and exhaustive nominal categories, or (2) rank order the applicability of categories to subjects, is desirable. A generalized ANOVA model is presented which allows the estimation of various reliability coefficients of interest for all classification tasks described. (Author)

ED106354

A GENERALIZED ANOVA MODEL FOR ESTIMATING THE RELIABILITY OF CATEGORICAL JUDGMENTS

John M. Enger and Douglas R. Whitney
University of Iowa

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Presented at the Annual Meeting of the
American Educational Researchers Association
Washington, D.C.
April, 1975

TM 004 470

ABSTRACT

A GENERALIZED ANOVA MODEL FOR ESTIMATING THE RELIABILITY OF CATEGORICAL JUDGMENTS

John M. Enger and Douglas R. Whitney

There are few existing or widely known measures of agreement applicable when data is nominal or categorical. Most such coefficients are applicable only when judges classify objects or subjects into a single category. A wider range of applications is desirable to include those where judges (1) place probabilities on subjects belonging to mutually exclusive and exhaustive nominal categories or (2) rank order the applicability of categories to subjects, or (3) assign weights to the appropriateness of the placement of subjects into each of a set of categories. A generalized ANOVA model is presented which allows the estimation of various reliability coefficients of interest for all assignment tasks described.

A GENERALIZED ANOVA MODEL FOR ESTIMATING THE RELIABILITY OF CATEGORICAL JUDGMENTS

Situations frequently arise in research activities which require that the experimenter obtain from judges estimates of the degree to which each of a set of objects belongs to each of a prescribed set of classes or categories. Conventionally, the degree of agreement among judges is presented as a measure of the reliability of the judgments. In the familiar case in which the ratings solicited from judges represent an amount of a single trait or characteristic, analysis of variance techniques have provided useful reliability indices (c.f.g., Ebel, 1951). The situations of concern in this paper, however, are of a different nature. Our concern is with situations in which each judge is required to simultaneously evaluate the degree to which an object possesses a specified set of traits or characteristics.

Four recent examples will serve to illustrate this kind of judgmental activity:

1. Pyrczak and Rasmussen (1973) asked two judges to classify each of 52 items on a standardized reading test into one of seven categories.
2. Board and Whitney (1972) asked six judges to classify each of 20 multiple-choice test items into one of five categories according to whether any of four poor item-writing practices had been used. (The fifth category was "no flaws.")
3. Robinson (1974) asked three judges to allocate each of 232 test items, from the Iowa Tests of Basic Skills, integer values of 0, 1 or 2 according to the degree to which each item required the use of three cognitive learning styles (relational, descriptive, categorical).

4. Enger (1975) asked 16 students in an educational psychology class to allocate each of 25 test items, from a standardized test in educational psychology, integer values of 0-9 according to the degree to which each item represented three major content areas of the course. Other assignment tasks investigated included those where students placed their probabilities as to the appropriateness of the placement of test items into three content areas and one in which students ranked the appropriateness of the placement of test items into three content areas.

Many procedures are available for the first situation (two judges, simple classification) and one has been extended to the second (more than two judges, simple classification). No well-known procedures are available for quantifying the degree of agreement among judges. This paper will

a. describe, illustrate, and compare two indices appropriate for simple classification,

b. develop a generalized analysis of variance for expressing the reliability of such judgments which is appropriate for all situations described above, and

c. illustrate the applications of the generalized technique to potential research situations.

AGREEMENT AND ASSOCIATION

For the simplest judgment situation (two judges, simple classification), there are a number of indices which express the strength of the relationship between judges' classifications. (That is, the degree to which the joint frequencies differ from those estimated from the products of the marginal

frequencies.) Some of these indices are the contingency coefficient (Guilford, 1956), the lambda coefficient and Guttman's lambda coefficient (Goodman and Krushal, 1954) which are useful as measures of association, but do not elicit values of agreement exclusively in the [1,-1] range and thus are unacceptable as indications of reliability. An index of agreement among judges should be unity if and only if all judges agree exactly on the assignment of all subjects. The expected value when there is no relationship between judges should be approximately zero.

INDICES OF AGREEMENT

Simple Classification, Two Judges

When two judges classify each of s objects into one of c mutually exclusive and exhaustive classes, the frequencies may be displayed as a c x c table. Cohen (1960) illustrated the situation with an example in which two judges (psychiatrists) placed subjects (patients) into three categories (1=schizophrenic, 2=neurotic, and 3=brain-damaged). Figure 1 illustrates frequencies from a second example in the same article.

Insert Figure 1 about here

Scott (1955) proposed a coefficient of intercoder agreement which involves the frequency of agreement between judges. Since some agreements would be expected to occur even under random classifications, Scott standardized the frequency of agreement by the frequency expected by chance. The latter was based on the squares of the average marginal frequencies. Specifically,

$$\pi = \frac{\sum_{j=1}^{c} f_{jj} - \sum_{j=1}^{c} \left(\frac{f_{1j} + f_{2j}}{2} \right)^2}{N - \sum_{j=1}^{c} \left(\frac{f_{1j} + f_{2j}}{2} \right)^2}, \qquad (1)$$

where f_{jj} is the relative frequency with which judges agreed on placement in category j, f_{1j} and f_{2j} are the relative frequencies with which the two judges used category j, and n is the number of objects classified; N represents the total number of observations. Using the data from Figure 1,

$$\pi = \frac{(88+40+12) - \left(\frac{100^2 + 60^2 + 30^2}{200} \right)}{200 - \left(\frac{110^2 + 60^2 + 30^2}{200} \right)} = .487.$$

Note that to obtain the relative frequency of agreement expected by chance, Scott used the average marginal frequency for each category. In the event that the marginal frequencies differ between judges, π cannot reach unity. The expected value of π is near zero under the hypothesis of independent (random) classifications.

A later coefficient, kappa (k), described by Cohen (1960) differs from π only in the manner in which the relative frequency of agreement expected by chance is computed. Cohen used the sum of cross products of the marginal relative frequencies instead of the squared average marginals as in π. Cohen's coefficient is

$$k = \frac{\sum_{j=1}^{c} f_{jj} - \sum_{j=1}^{c} \frac{f_{1j} f_{2j}}{N}}{N - \sum_{j=1}^{c} \frac{f_{1j} f_{2j}}{N}} . \qquad (2)$$

Using the data in Figure 1,

$$k = \frac{(88+40+12) - (60+18+4)}{200 - (60+18+4)} = .492 .$$

It is easy to establish that, for any set of data, $k \geq \pi$ with equality holding only when the marginal frequencies for each category are identical for both judges. The expected value of k under random classification is near zero (Everitt, 1968).

Simple Classification, Two or More Judges

To extend Scott's π coefficient for more than two judges, it is simply necessary to use the squares of the c category marginal frequencies to obtain the adjustment for "chance" agreements. Let f_{ijk} be 1 if object k was classified in category j by judge i and 0 if not. The expected relative frequency of agreement due to chance is $\sum_{j=1}^{c} f_{\cdot j \cdot}^{2} / (rs)^{2}$ where the dot denotes a summation over the deleted subscript. The extension of π is then

$$\pi = \frac{\frac{\sum_{j=1}^{c} \sum_{k=1}^{s} f_{\cdot jk}^{2} - rs}{rs(r-1)} - \frac{\sum_{j=1}^{c} f_{\cdot j \cdot}^{2}}{(rs)^{2}}}{1 - \frac{\sum_{j=1}^{c} f_{\cdot j \cdot}^{2}}{(rs)^{2}}} . \qquad (3)$$

In a recent article, Fleiss (1971) proposed an extension of k which is identical to equation (3). This extension is more appropriately considered an excension of π because of the use of pooled or averaged marginal frequencies to obtain the frequency of expected agreements due to chance.

Figure 2 presents data obtained by Board and Whitney by having six judges classify test items into five categories. Using this data, the extension of π takes the following value:

$$\pi = \frac{\frac{(6^2+0^2+0^2+\cdots+0^2)-(6)(20)}{(6)(20)(5)} - \frac{(20^2+18^2+25^2+14^2+43^2)}{[(6)(20)]^2}}{1 - \frac{(20^2+18^2+25^2+14^2+43^2)}{[(6)(20)]^2}}$$

$$= \frac{\frac{542-120}{600} - \frac{3394}{14400}}{1 - \frac{3394}{14400}} = .612 .$$

In order to extend k properly, one should use the sum of the cross products of pairs of marginal frequencies to obtain the expected agreements by chance (Light, 1971). Thus, the extended k coefficient would be

$$\frac{\frac{\sum_{j=1}^{c}\sum_{k=1}^{s} f_{\cdot jk}^2 - rs}{rs(r-1)} - \frac{2\sum_{i=1}^{r-1}\sum_{i'=i+1}^{r}\left(\sum_{j=1}^{c} f_{ij\cdot} f_{i'j\cdot}\right)}{rs^2(r-1)}}{1 - \frac{2\sum_{i=1}^{r-1}\sum_{i'=i+1}^{r}\left(\sum_{j=1}^{c} f_{ij\cdot} f_{i'j\cdot}\right)}{rs^2(r-1)}} . \qquad (4)$$

This equation would be very awkward as a computing method, but it can be easily verified that Equation 5 is algebraically equivalent.

$$k = \frac{\dfrac{\sum_{j=1}^{c}\sum_{k=1}^{s} f_{\cdot jk}^{2} - rs}{rs(r-1)} - \dfrac{\sum_{j=1}^{c} f_{\cdot j\cdot}^{2} - \sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij\cdot}^{2}}{rs^{2}(r-1)}}{1 - \dfrac{\sum_{j=1}^{c} f_{\cdot j\cdot}^{2} - \sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij\cdot}^{2}}{rs^{2}(r-1)}} \qquad (5)$$

Using the data from Figure 2, the k coefficient takes the value

$$k = \frac{\dfrac{542 - 120}{600} - \dfrac{3394 - 610}{12000}}{1 - \dfrac{3394 - 610}{12000}} = .614 .$$

Thus, as for the simpler case, $k \geq \Pi$ although the difference for this data is negligible. The difference between values increases as the judges' category marginal frequencies become more disparate.

General Methods

Assume that each of r judges assigns some value x_{ijk} to represent the proximity or resemblance of object k to category j. These data may be displayed in an $r \times c \times s$ matrix with each cell containing a single observation. Following the usual ANOVA procedures, the total sum of squares (SS_{TOT}) can be partitioned as:

$$SS_{TOT} = SS_{R} + SS_{C} + SS_{S} + SS_{RC} + SS_{SC} + SS_{RS} + SS_{RCS} \qquad (6)$$

Each of these sums of squares takes on a meaning directly related to the concept of the reliability of judgments:

1. Sum of squares for raters or judges (SS_R) reflects the differences among judges of the average values assigned across objects and categories. It is analogous to the differences in raters' "levels" in a univariate rating task and would usually be considered as "error."
2. Sum of squares for categories (SS_C) reflects the differences among categories in the average values assigned across raters and objects. For a suffieciently large number of raters, this component simply describes the sample of objects. Like the sum of squares for items in the Hoyt (1941) procedure for estimating test reliability, this component represents neither "true" nor "error" variance.
3. Sum of squares for objects or subjects (SS_S) reflects the differences among objects in the average values assigned across judges and categories. It would usually represent "error" variance.
4. Sum of squares for judges-by-categories (SS_{RC}) reflects the differences among average values assigned by judges to each category. Again, this component would usually be considered to be "error" variance.
5. Sum of squares for objects-by-categories (SS_{SC}) reflects differences among average values assigned to each object-category combination. This, presumably, reflects "true" variance, since it is assumed that most objects "fit" one category better than the others.
6. Sum of squares for judges-by-objects (SS_{RS}) reflects differences among average values assigned by judges to each object. Again, this source usually reflects "error" variance.
7. Sum of squares interaction (SS_{RCS}) reflects residual variance--also an "error" component in reliability considerations.

Thus, if we follow the usual procedure for estimating the average reliability for a single rater, we would use the general form

$$\text{reliability} = \frac{MS_{subj.} - MS_{error}}{MS_{subj.} + (r-1)\,MS_{error}} . \qquad (7)$$

By taking $MS_{subj.}$ to be MS_{SC} and varying the definition of MS_{error}, we can obtain an array of reliability coefficients reflecting desired sources of "error." As an example, a coefficient which would reflect all sources of "error" would be

$$\text{reliability} = \frac{MS_{CS} - \left\{ \frac{MS_R}{(c-1)(s-1)} + \frac{MS_S}{(r-1)(c-1)} + \frac{MS_{RC}}{(s-1)} + \frac{MS_{RS}}{(c-1)} + MS_{RCS} \right\}}{MS_{CS} + (r-1)\left\{ \frac{MS_R}{(c-1)(s-1)} + \frac{MS_S}{(r-1)(c-1)} + \frac{MS_{RC}}{(s-1)} + \frac{MS_{RS}}{(c-1)} + MS_{RCS} \right\}} \qquad (8)$$

Note that MS_C is never included, since it reflects neither "true" nor "error" variance.

This generalized ANOVA approach makes possible the estimation of the reliability of categorical judgments for nearly any conceivable situation. In addition, however, it also identifies and estimates the consequences of the various sources of "error" variance. There may also be cases in which formal tests of hypotheses concerning these effects are desired and they would be possible using this framework. Finally, this approach allows for a simple solution to certain problems involving missing data. In most cases, the relevant terms will still be estimable even for less-than-complete data.

Application of the General Procedure to Simple Classification

When judges are instructed to classify each object into a single category, of course all x_{ijk} values become 0's or 1's. In this situation, the constraints imposed cause SS_R, SS_S and SS_{RS} to be zero and SS_{TOT} to be $\frac{c-1}{c}(rs)$. Under this condition, Equation 8 simplifies to

$$\pi = \frac{MS_{CS} - \left\{ \frac{MS_{CR}}{s-1} + MS_{RCS} \right\}}{MS_{CS} + (r-1)\left\{ \frac{MS_{CR}}{s-1} + MS_{RCS} \right\}}$$

$$= \frac{SS_{CS} - (SS_{CR} + SS_{RCS})/(r-1)}{SS_{CS} + SS_{CR} + SS_{RCS}} . \qquad (9)$$

It can be readily verified that this equation yields a value exactly equal to that resulting from Equation 3. Thus, this coefficient appears to have a very solid analytical basis and probably represents the most useful approach for simple classification problems.

As an alternative, one could consider the average correlation for each category between assignments to the category, across all possible pairs of judges, averaged across all categories. Such a coefficient would have the desirable feature of being interpretable as an "expected" correlation between a pair of judges for any category. The numerator of the average of all $\frac{1}{2}cr(r-1)$ such correlations is proportional to $SS_{CS}-SS_{RCS}/(r-1)$. That suggests the use of the coefficient

$$r_{pooled} = \frac{MS_{CS} - MS_{RCS}}{MS_{CS} + (r-1)MS_{RCS}}$$

$$= \frac{SS_{CS} - SS_{RCS}/(r-1)}{SS_{CS} + SS_{RCS}} . \qquad (10)$$

Alternatively, this coefficient may be derived by analyzing each category separately into three components ($SS_{objects}$, SS_{judges} and $SS_{interaction}$) and pooling (summing) these terms across categories before applying the Hoyt (1941) procedure to the pooled sums of squares.

This coefficient will be larger than that resulting from Equation 9--a finding consistent with the fact that the difference among category marginal frequencies for judges is ignored in the computations. This coefficient, however, would clearly have an expected value of zero under the hypothesis of independent classifications.

It may be of interest to note that the extended k coefficient of Equation 5 can be expressed as

$$k = \frac{MS_{CS} - MS_{RCS}}{MS_{CS} + (r-1)\left\{MS_{RCS} + \frac{r}{(r-1)} \cdot \frac{MS_{CR}}{(s-1)}\right\}}$$

$$= \frac{SS_{CS} - SS_{RCS}/(r-1)}{SS_{CS} + SS_{RCS} + \frac{r}{(r-1)} \cdot SS_{CR}} \quad . \qquad (11)$$

The interpretation, in terms of reliability components, is not clear, at present, for this coefficient. It is, however, interesting to note that the numerator is identical with that for Equation 10 while the denominator is similar to that for Equation 9.

ILLUSTRATION OF GENERAL METHODS

Figure 3 displays the data (Enger, 1975) for 3 educational psychology students who assigned integer values 0-9 to 10 test items to reflect their relationship to 3 content areas. The relevant ANOVA terms are:

SS_{TOT}=957.40

SS_R = 7.47

SS_C =207.80

SS_S = 18.40

SS_{CS} =416.20

SS_{CR} = 56.93

SS_{RS} = 20.53

SS_{RCS}=230.07

Thus, the comprehensive coefficient of Equation 8 has the value

$$\text{reliability} = \frac{23.12 - (0.21+0.51+1.58+0.57+6.39)}{23.12 - 2(0.21+0.51+1.58+0.57+6.39)}$$

$$= \frac{23.12 - 9.26}{23.12 + 18.52}$$

$$= .333 .$$

Other coefficients of interest are

$$\Pi = \frac{MS_{CS} - \left\{ \frac{MS_{CR}}{(r-1)} + MS_{RCS} \right\}}{MS_{CS} + (r-1)\left\{ \frac{MS_{CR}}{(r-1)} + MS_{RCS} \right\}} = \frac{23.12 - 7.97}{23.12 + 15.94} = .388 ,$$

$$K = \frac{MS_{CS} - MS_{RCS}}{MS_{CS} + (r-1)MS_{RCS} + \frac{r}{(s-1)} MS_{CR}} = \frac{23.12 - 6.39}{23.12 + 12.78 + 4.74} = .412 ,$$

$$r_{pooled} = \frac{MS_{CS} - MS_{RCS}}{MS_{CS} + (r-1)\, MS_{RCS}} = \frac{23.12 - 6.39}{23.12 + 12.78} = .466 .$$

In order to evaluate the effects of restricting judges to a simple classification (rather than the unrestrained weights), Enger forced a post-hoc

classification based on the highest category weight for each judge. Recomputed using the classification "weights," he obtained

$$\pi = .214,$$

$$k = .240, \text{ and}$$

$$r_{pooled} = .266.$$

The dramatic decline in values for all coefficients probably reflects the loss of information due to the restraints of classification. This suggests that a greater number of judges are required to attain reliabilities for classification tasks equal to those for unconstrained weights. In this example, it would require about 2.4 times as many judges to achieve an r_{pooled} value for classification tasks equal to that for the weighting task.

SUMMARY

Generalized procedures have been described to facilitate the estimation of reliability coefficients for a wide variety of classification tasks and related multiple-category judgment decisions. This approach provides a means for better identifying the sources of error variance and for testing hypotheses concerning these sources.

References

Board, C. & Whitney, D. R. The effect of selected poor item-writing practices on test difficulty, reliability and validity. Journal of Educational Measurement, 1972, 9, 225-233.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 6, 407-424.

Enger, J. M. Assignment tasks, category types and use of students in the content validation of standardized examinations. Unpublished doctoral dissertation. University of Iowa, 1975.

Everitt, B. S. Moments of the statistics kappa and weighted kappa. British Journal of Mathematical and Statistical Psychology, 1968, 21, 97-103.

Fleiss, J. L. Measuring nominal scale agreement among many raters. Psychological Bulletin, 1971, 76, 378-382.

Goodman, L. A. & Kruskal, W. H. Measures of association for cross classifications. Journal of the American Statistical Association, 1954, 49, 732-764.

Guilford, J. P. Fundamental statistics in psychology and education. (2nd ed.) New York: McGraw-Hill, 1950.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Light, R. J. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin, 1971, 76, 365-377.

Pyrczak, F. & Rasumssen, M. Subjective analysis of the skills measured by selected reading tests designed for use in high school. Presented at the annual meeting of the California Educational Research Association, Los Angeles, November, 1973.

Robinson, J. An investigation of the relationship between cognitive style and school learning via a multitrait multimethod matrix method. Unpublished doctoral dissertation. University of Iowa, 1974.

Scott, W. A. Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 1955, 19, 321-325.

Figure 1

Classification of 100 objects

to 3 nominal categories by 2 judges

		Judge 1			
		Category 1	Category 2	Category 3	
Judge 2	Category 1	88	14	18	120
	Category 2	10	40	10	60
	Category 3	2	6	12	20
		100	60	40	N=200

Figure 2

Assignment of 20 test items

to 5 categories by 6 raters

Assignment of Items to Categories by Raters

Items	Categories				
	1	2	3	4	5
1	6	0	0	0	0
2	0	0	6	0	0
3	0	6	0	0	0
4	0	0	0	3	3
5	2	0	0	0	4
6	0	0	0	1	5
7	0	0	2	1	3
8	6	0	0	0	0
9	0	0	0	1	5
10	0	0	4	1	1
11	0	0	5	0	1
12	0	0	0	0	6
13	0	0	0	1	5
14	0	0	5	0	1
15	0	0	0	6	0
16	0	0	0	4	2
17	0	0	0	0	6
18	0	3	2	0	1
19	0	4	0	2	0
20	6	0	0	0	0
Totals	20	18	25	14	43

Number of Items Assigned to Each Category by Rater

Raters	Categories				
	1	2	3	4	5
1	3	2	2	4	9
2	3	3	5	5	4
3	4	3	3	1	9
4	3	4	5	1	7
5	3	3	5	2	7
6	4	3	5	1	7
Totals	20	18	25	14	43

Figure 3

Weights assigned to 10 test items for 3 content categories by 3 judges

	Judge 1			Judge 2			Judge 3		
	Category 1	Category 2	Category 3	Category 1	Category 2	Category 3	Category 1	Category 2	Category 3
Item 1	7	0	2	6	4	2	8	4	4
Item 2	0	0	9	4	4	4	9	0	3
Item 3	5	7	0	8	2	0	7	4	0
Item 4	6	0	5	7	5	4	8	0	6
Item 5	4	8	0	4	8	0	9	6	0
Item 6	0	9	0	0	9	0	5	9	0
Item 7	9	0	0	8	1	0	8	2	0
Item 8	6	5	3	8	1	0	4	8	0
Item 9	5	0	6	6	2	4	3	8	2
Item 10	2	0	8	8	1	0	9	0	0

20